DOCUMENT RESUME

ED 458 241 TM 033 402

AUTHOR Lee, Yong-Won; Golub-Smith, Marna; Payton, Carmen; Carey, Jill

TITLE The Score Reliability of the Test of Spoken English (TSE) from the Generalizability Theory Perspective: Validating the Current Procedure.

PUB DATE 2001-04-00

NOTE 17p.; Paper presented at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001).

PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)

EDRS PRICE EDRS Price MF01 Plus Postage. PC Not Available from EDRS.

DESCRIPTORS *Adults; English (Second Language); Estimation (Mathematics); *Generalizability Theory; Psychometrics; *Reliability; *Scores; Second Language Learning; *Validity

IDENTIFIERS *Test of Spoken English

ABSTRACT

This study investigated the validity of the current reliability estimation procedure for the Test of Spoken English (TSE), a tape-mediated semi-performance test of 12 speaking tasks, from the perspective of generalizability theory and examined the feasibility of shortening the test without compromising the psychometric quality of the test. Data were ratings from 2 TSE administrations: (1) ratings assigned by 80 raters for 1,766 test takers; and (2) ratings given by 81 raters for 1,563 examinees. The results provide useful information to the TSE program about the minimal number of items required for a shorter version of the TSE. The current TSE score reliability estimation procedure seems to be working well for the full 12-item test, providing more conservative estimates of test score reliability than the other coefficients examined. From a generalizability theory perspective, it appeared that a high score reliability could still be maintained if the test were reduced to 9 to 10 items. (Contains 3 tables, 4 figures, and 11 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made from the original document.

ED 458 241

The Score Reliability of the Test of Spoken English (TSE) from the Generalizability Theory Perspective: Validating the Current Procedure

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY HAS BEEN GRANTED BY

Y-W. Lee

2A TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Yong-Won Lee

Marna Golub-Smith

Carmen Payton

and

Jill Carey

Division of Statistics and Research

Educational Testing Service

Rosedale Road Princeton

NJ 08541

Paper presented at the annual conference of American Educational Research Association (AERA) held in Seattle, Washington between April 10-14, 2001.

TM033402

BEST COPY AVAILABLE

Abstract

The present study investigates the validity of the current reliability estimation procedure for the Test of Spoken English (TSE) from the perspective of *generalizability theory* and examines the feasibility of shortening the test without compromising the psychometric quality of the test. The results of the study provide useful information to the TSE program about the minimal number of items required for a shorter version of TSE.

The Score Reliability of the Test of Spoken English (TSE) from the G-theory Perspective: Validating the Current Procedure

I. Introduction

In performance assessments, it is usually necessary to grapple with two or more facets simultaneously – especially tasks and raters. In the classical test theory framework, these measurement facets can be examined only one at a time in relation to the object of measurement (persons), as exemplified by inter-rater or internal consistency reliabilities. Generalizability (G) theoy liberalizes the CTT by providing a conceptual framework and methodology for disentangling multiple sources of error simultaneously from the undifferentiated error (*E*) term (Brennan, 2000b). The present study has two major goals: (1) To investigate the validity of the current reliability estimation procedure used for the Test of Spoken English (TSE) from the G-theory perspective; (2) to examine the feasibility of shortening the test without compromising the psychometric quality of the test.

II. TSE score reliability from the generalizability theory framework

Generalizability (G) theory (Brennan, 1992; Brennan, 2000) is a comprehensive conceptual and methodological framework that can analyze multiple measurement facets including the object of measurement (usually examinees) simultaneously that are involved in the assessment system. The *G*-theory also enables us to evaluate the relative importance of various sources of measurement error, investigate the projected test score reliabilities for different levels of the measurement facets (*e.g.*, different test length, different number of raters), and interpret the test score reliability both from norm-referenced and criterion-referenced frameworks (*e.g.*, a *generalizability coefficient* (G) and a *dependability index* (Φ)).

G-theory is well-suited especially for analyzing scores from performance assessments, where both tasks and raters are involved as measurement facets. The Test of Spoken English (TSE) is a tape-mediated semi-performance test that consists of 12 speaking tasks rated by 2 independent raters. Inter-rater and internal consistency reliabilities are currently reported to provide information about the psychometric quality of ratings, but do not provide a test score reliability index that take into account the two major sources errors (rating and tasks) simultaneously. For this reason, a test score reliability coefficient was formulated that incorporates the two sources errors in a single index. However, this score reliability coefficient also can be seen as a simplified version of a *generalizability* (*G*) coefficient in the G-theory framework.

The current TSE reliability estimation procedures can be discussed in relation to the G-theory reliability indices, as summarized in Table 1.

Current procedure: inter-rater, test form, and test score reliabilities

Interrater reliability (α_R). The inter-rater reliability provides information about rating consistency. In TSE, however, a different pair of raters is randomly assigned to different examinees for rating examinee response tapes, and, thus, it is not feasible to compute inter-rater reliability by using raters as units of analyses. Instead, alpha (α_R) is computed based on the first and second ratings, as shown in equation 1:

$$\alpha_R = 2(1 - \frac{\sigma^2_{R1} + \sigma^2_{R2}}{\sigma^2_{R1+R2}}) \equiv \frac{\sigma^2_p}{\sigma^2_p + \frac{\sigma^2_{pr'}}{n_{r'}}}, \quad \text{..... (1)}$$

where R_1 and R_2 are the averaged ratings assigned to an examinee in the first and second ratings, respectively, (ETS, 1999). This formula treats the two rating scores as if they were items on a two-item test, with the total score being the sum of the two rating scores. The inter-rater reliability can be expressed in terms of the person (σ_p^2) and the person-and-rating interaction ($\sigma_{pr'}^2/n_{r'}$) components in a single-facet crossed design (p x R'), as in Equation 1. The Cronbach's (1951) coefficient α equals a *generalizability* (G) coefficient in a single-facet crossed design (Brennan, 1992; Suen, 1990).

Test form reliability (α_I). The test form reliability (α_I) is computed by using Cronbach's (1951) coefficient alpha. The test form reliability provides information about the internal consistency of a TSE form or the extent to which the items in a TSE test form are assessing the same construct.

$$\alpha_I = \frac{k}{k-1}(1 - \frac{\sum_{i=1}^{k}\sigma^2_i}{\sigma^2_I}) \equiv \frac{\sigma^2_p}{\sigma^2_p + \frac{\sigma^2_{pi}}{n_i}}, \quad \text{..... (2)}$$

where k equals 12, the total number of items in the test, σ_i^2 is the variance of item i, and σ_I^2 is the variance of the sum of the 12 items (ETS, 1999). The internal consistency reliability can be expressed in terms of the person (σ_p^2) and the person-and-item interaction ($\sigma_{pi'}^2/n_i$) components, as in Equation 2, in a single-facet crossed design (p x I). It should be noted, however, that the inter-rater and test form reliability coefficients are expressed on the total score metric in the alpha, whereas they are expressed on the mean score metric in the G-theory coefficient.

TSE Score reliability (α_T). The TSE score reliability coefficient (α_S) is currently computed by using a short-cut formula that combines the inter-rater (α_R) and the test form reliabilities (α_I). As shown in equation 3, it is similar in form to the G-coefficient (that will be explained later in the next section), except that the three-way interaction term ($\sigma_{pr'i}^2/n_{r'}n_i$) is not included in the denominator as part of the error variance. The current procedure does not attempt to estimate all the main and interaction variance components as in G-theory analyses, but uses the inter-rater reliability (α_R) and the test form reliability (α_I) to estimate the test score reliability (α_S), as shown in equation 3:

$$\alpha_S^* = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{pr'}^{2*}}{n_{r'}} + \frac{\sigma_{pi}^2}{n_i}} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_p^2 \frac{1-\alpha_R^*}{\alpha_R^*} + \sigma_p^2 \frac{1-\alpha_I}{\alpha_I}} = \frac{\alpha_R^* \alpha_I}{\alpha_R^* + \alpha_I - \alpha_R^* \alpha_I} \quad \ldots (3)$$

* indicates that the asterisked components are estimated from the ratings before adjudication.

One important thing to mention is that, in order to prevent α_S from being an overestimate of the test score reliability, the current procedure reports the test score reliability (α_S^*) that uses the inter-rater reliability (α_R^*) estimated from the ratings before adjudication as a primary index, but provides the test score reliability (α_S) estimate based on the ratings after adjudication (α_R) for comparison as well. The inter-rater reliability is somewhat higher after adjudication.

Generalizability coefficient and dependability index

When multiple raters (r) and items (i) are involved in the assessment of examinee (p) performance, for instance, the most common G-study designs might be a fully-crossed design (p x r x i) and a partially-nested design with raters nested within examinees ((r:p) x i). These two designs, however, may not be appropriate in a strict sense in the context of the TSE, because neither a crossed nor a nested design is currently used for TSE examinees and raters. Although two raters' ratings are averaged to compute the final reported score for each examinee in TSE, the same rater pair is not systematically assigned to a block of examinees, as in a typical nested design. Instead, the rater pair is somewhat randomly decided for each examinee.

Nevertheless, it may be possible to apply a fully-crossed design with two random facets of "ratings" (not raters) and "items" (p x r' x i) by treating ratings (*i.e.*, the first and second ratings: r') as a random facet instead of raters (r). Because raters are strictly trained according to a scoring rubric and because their performance is continuously monitored for supplemental training, ratings and raters can be assumed to be drawn from a relatively homogeneous universe (Bachman, Lynch, & Mason, 1995; Lynch McNamara, 1998). In addition, it is also consistent with the logic behind the inter-rater reliability (α_R) computation in the current procedure, for the ratings (not raters) to be used as a unit of analyses to obtain the inter-rater reliability estimate.

Under such crossed-design assumptions, all of the 7 main and interaction variance components (σ_p^2, $\sigma_{r'}^2$, σ_i^2, $\sigma_{pr'}^2/n_{r'}$, σ_{pi}^2/n_i, $\sigma_{ir'}^2/n_i n_{r'}$, $\sigma_{pr'i}^2/n_{r'} n_i$) can be estimated, and two different kinds of score reliability equivalents can be computed, that is, a *generalizability coefficient* (G) and a *dependability index* (Φ).

First, the relative error term (δ) and the generalizability coefficient (G) can be defined as in equation 4, which can be interpreted as the error variance and the reliability index for norm-referenced score interpretation, respectively (Brennan, 1992; Suen, 1990).

$$G = \frac{\sigma_p^2}{\sigma_p^2 + \delta} = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{pr}^2}{n_r} + \frac{\sigma_{pi}^2}{n_i} + \frac{\sigma_{pri}^2}{n_r n_i}} \quad \text{.....(4)}$$

Second, the absolute error (Δ) and the dependability index (Φ) can be computed as in equation 5, which can be interpreted as the error variance and the score reliability index for criterion-referenced score interpretation, respectively. Main score categories in TSE are supposed to have somewhat qualitative, criterion-related meanings about test takers' speaking proficiency according to the rating scale. Especially when the cut-off is applied for admission and licensure purposes, the Φ coefficient would be more appropriate for TSE score interpretation.

$$\Phi = \frac{\sigma_p^2}{\sigma_p^2 + \Delta} = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_r^2}{n_r} + \frac{\sigma_i^2}{n_i} + \frac{\sigma_{pr}^2}{n_r} + \frac{\sigma_{pi}^2}{n_i} + \frac{\sigma_{ri}^2}{n_r n_i} + \frac{\sigma_{pri}^2}{n_r n_i}} \quad \text{.....(5)}$$

Another possible way to estimate a TSE score reliability is to use a partially-nested design with raters nested within examinees ((*r*:*p*) x *i*) mentioned previously, even though raters are not nested within examinees in a strict sense of nesting in TSE (Brennan, personal communication, 2001). The *G* and Φ coefficients obtained from this (R:*p*) x I *design* can be compared to those estimated from the *p* x R˚x I design to examine if it would be reasonable to treat ratings as a facet in TSE. The mathematical equations for the *G* and Φ coefficients for this nested design are also provided in Table 1.

III. Method

Data

Data used for this study are ratings from two TSE administrations: (1) ratings assigned by 80 raters over 12 items for 1766 test-takers who took a January 1999 TSE form and (2) ratings given by 81 raters for 1563 examinees who took an April 1999 TSE form. In TSE, 12 items are presented to elicit the examinee responses that are rated on a scale of 20 to 60 by two raters. When there is a summed rating discrepancy greater than or equal to120 over 12 items between the first and second ratings, a third rater re-rates the tape and the two closest ratings are used to compute the final reported score. The final score is an average of two final ratings that are reported in increments of 5 (e.g., 20, 25, … 55, 60).

Analyses

The computer program GENOVA (Crick & Brennan, 1983) was used to estimate the variance components for the main and interaction effects and *G* and Φ coefficients for

all the designs employed in this study. Separate analyses were done for January and April 1999 data.

First, two single-facet designs with rating (p x r') and item (p x i) as a random facet, respectively, were used to compute the inter-rater (α_R) and test form (α_T) reliabilities for a full-12 item test and also for shorter test lengths. Separate GENOVA analyses were done for ratings before and after adjudication for the p x r' design. The score reliabilities (α_S*, α_S) were obtained by combining the inter-rater (α_R*, α_R) and test form reliability (α_T) estimates from the GENOVA runs in the same way as in the current procedures.

Second, a two-facet crossed design (p x r' x i) with rating and item as random facets was employed as a G-study design to estimate the variance components for the main and interaction effects and G and Φ coefficients. A two-facet nested design with raters nested within examinees ((r:p) x i) was also used to estimate G and Φ coefficients for comparison. Multiple D-studies were carried out by varying the number of items from 1 to 12 in both designs, and the G and Φ coefficients from both designs were compared to examine if both designs actually produce very close estimates.

Third, the inter-rater, the test form, and the test score reliability estimates were plotted with the G and Φ coefficients that were estimated in the p x R' x I design for different test lengths. These estimates were compared to each other separately for January and April 1999 data.

Fourth, the dependability indices, Φ(λ), for different cut-off scores (Brennan & Kane, 1977) were computed for differing test lengths for the p x r' x i design separately for the two different TSE forms.

Results

For both the January and April 1999 administration data, very similar results were obtained. First, the current procedure has yielded an α_S coefficient of .90 and an α_S* of .89, which is lower than a G coefficient (.91) for the full 12-item test for January and April 1999 forms, as shown in Table 2. This can be partially explained by the fact that the error variance ($\sigma_{pr'}^2/n_{r'} + \sigma_{pi}^2/n_i$) associated with α_S was somewhat larger than the relative error for the G-coefficient ($\sigma_{pr'}^2/n_{r'} + \sigma_{pi}^2/n_i + \sigma_{pr'i}^2/n_{r'}n_i$). This may seem counter-intuitive, because the relative error for G includes an additional three-way interaction variance component.

However, it should be noted that both the person-and-rating ($\sigma_{pr'}^2/n_{r'}$) and the person-and-item (σ_{pi}^2/n_i) interaction variance components for α_S are estimated separately from the p x R' and p x I designs, respectively. It may be that the person-rating-item interaction variance ($\sigma_{pr'i}^2/n_{r'}n_i$) is somehow absorbed into the person-and-rating interaction ($\sigma_{pr'}^2/n_{r'}$) and the person-and-item interaction variance (σ_{pi}^2/n_i) in each of the p x I and p x R' designs. For this reason, if these $\sigma_{pr'}^2/n_{r'}$ and σ_{pi}^2/n_i components are

added to form an error term for α_S, it may have the similar impact as the three-way interaction variance ($\sigma_{pr'i}^2/n_{r'}n_i$) is added twice to the relative error in the p x R' x I design.

Second, the G- and Φ coefficients obtained from the p x R' x I and ((R:p) x I) designs were very close (almost identical if rounded at the second decimal point) for different test lengths. The same pattern was observed both for January and April 1999 forms, indicating that it may be reasonable to treat ratings (instead of raters) as a measurement facet in TSE in computing the test score reliability within the G-theory framework.

Third, the G coefficient was slightly higher than the Φ coefficient as expected in the p x R' x I design, but both of them were very close in size for different test lengths because each of the rating, item, and rating-and-item interaction variance components was very small. Both α_S* and G coefficients increased, as the number of item increased, as shown in Figures 1 and 2, but also the difference between the two became larger gradually. Interestingly, the difference between the α_S and G coefficients was much smaller for different test lengths than the difference between α_S* and G, indicating that the α_S might be a better estimate of a G coefficient.

Fourth, the Φ(λ) coefficient was the lowest (.9122) at the reported score category of 45 that is very close to the mean score (45.33) for the full 12-item test and become larger and larger as the score category is further away from 45 toward the extreme scores. The same pattern was observed for different test lengths both for January and April 1999 data, and the increase in Φ(λ) was the most sensitive to the increase in the number of items at the score category of 45 that is closest to the examinee TSE score mean.

Conclusions

The current TSE score reliability estimation procedure seems to be working well for the full 12-item test, providing more conservative estimates of test score reliability than the G and Φ coefficients. In particular, the test score reliability (α_S) computed based on ratings after adjudication has produced an estimate much closer to the G coefficient than the one (α_S*) based on ratings before adjudication. It was also found out that a high score reliability (about .90) could be still maintained even after the test is reduced to a 9-10 item test from the G-theory perspective.

References

Bachman, L. F., Lynch, B. K. & Mason, M. (1995). Investigating variability in tasks and rater judgment in a performance test of foreign speaking. *Language Testing, 12*, 238-257.

Brennan, R. L. (1992). *Elements of generalizability theory*. Iowa City, IA: ACT.

Brennan, R. L. (2000a). (Mis)conceptions about generalizability. *Educational Measurement: Issues and Practice, 19(1)*, 5-10.

Brennan, R. L. (2000b). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement, 24(4)*, 339-353.

Brennan, R. L. (February, 2001). A pre-conference workshop on multivariate generalizability theory. The 23rd Language Testing Research Colloquium (LTRC) conference, St. Louis, MO.

Brennan, R. L. & Kane, M. T. (1977). An index of dependability of mastery tests. *Journal of Educational Measurement, 14*, 277-89.

Crick, J. E. & Brennan, R. (1983). GENOVA: A generalized analysis of variance system. Iowa city, IA: ACT.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Educational Testing Service. (1999). *Test of Spoken English test analysis* (SR-99-59). Unpublished statistical report. Princeton, NJ: ETS.

Lynch, B. K. & McNamara, T. F. (1998). Using g-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing 15* (2), 158-80.

Suen, H. K. (1990). *Principles of test theories*. Hillsdale, NJ: Lawrence Erlbaum.

Table 1. TSE score reliablities indices from classical and g-theory perspectives

---

A. The Current Procedure

(1) Inter-rater (rating) reliability: $$\alpha_R^* = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{pr^*}^2}{n_r}}$$

(2) Test form reliability (internal consistency)

$$\alpha_I = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{pi}^2}{n_i}}$$

(3) Score reliability:

$$\alpha_S^* = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{pi}^2}{n_i} + \frac{\sigma_{pr^*}^2}{n_r}} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_p^2 \frac{1-\alpha_R^*}{\alpha_R^*} + \sigma_p^2 \frac{1-\alpha_I}{\alpha_I}} = \frac{\alpha_R^* \alpha_I}{\alpha_R^* + \alpha_I - \alpha_R^* \alpha_I}$$

B. G-theory analyses [p x R' x I design]

(4) Generalizability Coefficient: $$G = \frac{\sigma_p^2}{\sigma_p^2 + \delta} = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{pr^*}^2}{n_{r^*}} + \frac{\sigma_{pi}^2}{n_i} + \frac{\sigma_{pr^*i}^2}{n_{r^*} n_i}}$$

(5) Dependability Index: $\Phi = \frac{\sigma_p^2}{\sigma_p^2 + \Delta} = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{r^*}^2}{n_{r^*}} + \frac{\sigma_i^2}{n_i} + \frac{\sigma_{pr^*}^2}{n_{r^*}} + \frac{\sigma_{pi}^2}{n_i} + \frac{\sigma_{r^*i}^2}{n_{r^*} n_i} + \frac{\sigma_{pr^*i}^2}{n_{r^*} n_i}}$

C. G-theory analyses [(R:p) x I design]

(6) Generalizability Coefficient: $$G = \frac{\sigma_p^2}{\sigma_p^2 + \delta} = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{r:p}^2}{n_{r^*}} + \frac{\sigma_{pi}^2}{n_i} + \frac{\sigma_{ri:p}^2}{n_{r^*} n_i}}$$

(7) Dependability Index: $\Phi = \frac{\sigma_p^2}{\sigma_p^2 + \Delta} = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{r:p}^2}{n_{r^*}} + \frac{\sigma_i^2}{n_i} + \frac{\sigma_{pi}^2}{n_i} + \frac{\sigma_{ri:p}^2}{n_{r^*} n_i}}$

Note: * indicates that the asterisked components are estimated based on ratings before adjudication.

Table 2. Variance components from D-studies for *p* x R', p x I, and *p* x R' x I designs

(a) *p* x R' ($n_{r'}$ =2) and p x I designs (n_i=12)

Effect	January 1999 Variance Components AA	(BA)	April 1999 Variance Components AA	(BA)
p	34.6051	(34.2476)	36.4471	(36.1476)
R'	0.3366	(0.0738)	0.0095	(0.0136)
*p*R'	2.6034	(3.1582)	2.8159	(3.3610)
α_R	.93	(.91)	.93	(.91)
p	33.0142		34.8139	
I	0.0260		0.0563	
*p*I	1.0041		1.1155	
α_T	.97		.97	
E (*p*R'+ *p*I)	3.6075	(4.1623)	3.9314	(4.4765)
α_S	.90	(.89)	.90	(.89)

Note: AA – after adjudication; BA – before adjudication

(b) *p* x R' x I design($n_{r'}$ =2; n_i=12)

Effect	January 1999 Variance Components	Total Variance (%)	April 1999 Variance Components	Total Variance (%)
p	32.5421	91.2	34.1713	90.8
R'	0.0321	0.09	0.0106	0.03
I	0.0261	0.07	0.0564	0.15
*p*R'	2.0712	5.81	2.2876	6.08
*p*I	0.5240	1.47	0.6239	1.66
R'I	0.0000	0	0.0001	0
*p*R'I	0.4817	1.35	0.4936	1.31
δ	3.0769		2.7812	
Δ	3.1351		3.4721	
G	.91		.91	
Φ	.91		.91	
	3.1351		3.4721	

Table 3. Comparison of G and Φ coefficient estimates from the *p* x R' x I and (R:*p*) x I designs

	January 1999				April 1999			
No. of	*p* x R' x I		(R:*p*) x I		*p* x R' x I		(R:*p*) x I	
Items	G	Φ	G	Φ	G	Φ	G	Φ
1	.69711	.69199	.69642	.69178	.68523	.67591	.68502	.67584
2	.80059	.79689	.79980	.79673	.79167	.78531	.79142	.78526
3	.84227	.83931	.84144	.83917	.83490	.83010	.83464	.83006
4	.86478	.86225	.86393	.86213	.85833	.85447	.85807	.85443
5	.87888	.87663	.87801	.87653	.87304	.86979	.87277	.86975
6	.88853	.88649	.88765	.88639	.88312	.88031	.88285	.88028
7	.89556	.89367	.89467	.89357	.89047	.88798	.89019	.88795
8	.90090	.89912	.90001	.89903	.89606	.89383	.89578	.89380
9	.90510	.90342	.90421	.90333	.90046	.89842	.90018	.89840
10	.90849	.90688	.90759	.90680	.90401	.90214	.90373	.90211
11	.91128	.90973	.91038	.90965	.90693	.90520	.90665	.90517
12	.91362	.91213	.91272	.91205	.90938	.90776	.90910	.90774

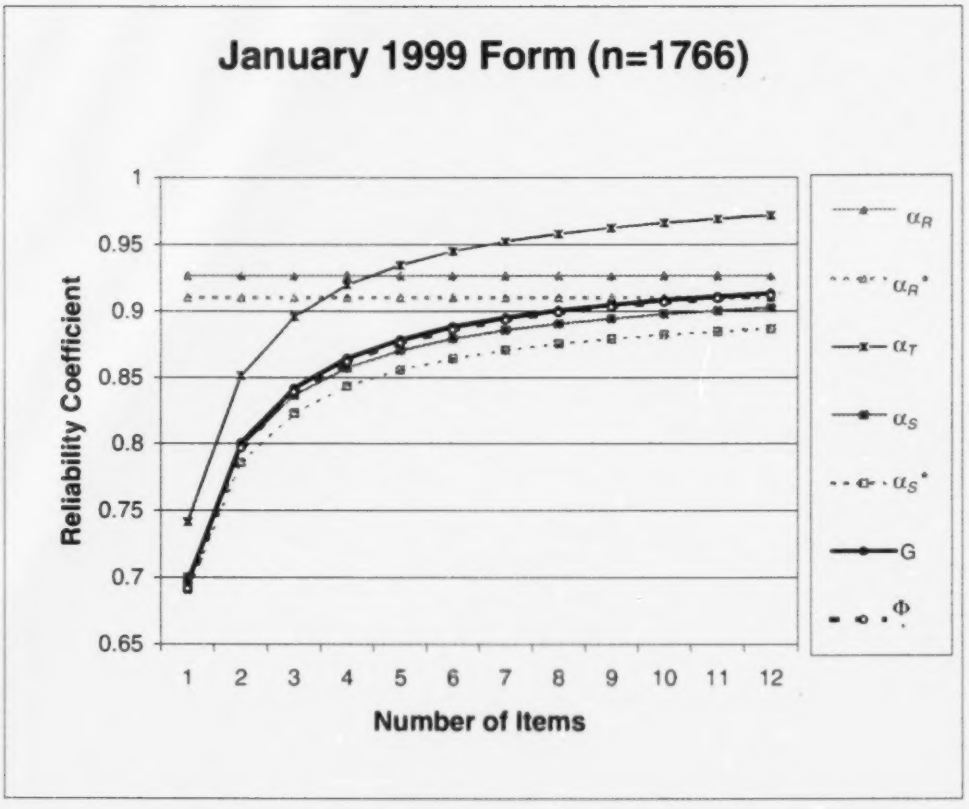

Figure 1. Inter-rater, test form, test score reliabilities, generalizability coefficient, and dependability index for different test lengths for January 1999 form

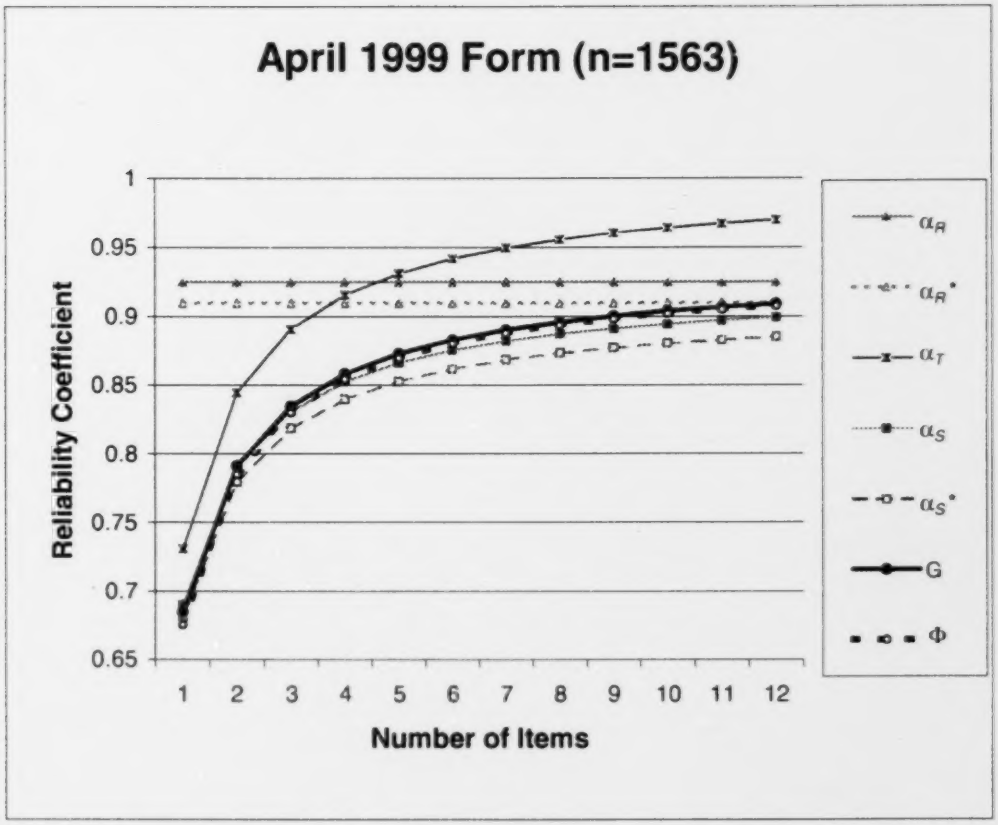

Figure 2. Inter-rater, test form, test score reliabilities, generalizability coefficient, and dependability index for different test lengths for April 1999 form

No. of Items	Different Cutoff Score Points									
	20	25	30	35	40	45	50	55	60	Mean (45.33)
1	.9790	.9685	0.9486	.9055	.8070	.6903	.7885	.8967	.9448	.6895
2	.9878	.9817	0.9699	.9437	.8798	.7964	.8672	.9382	.9676	.7958
3	.9908	.9862	0.9772	.9571	.9070	.8391	.8969	.9529	.9755	.8387
4	.9924	.9885	0.9809	.9640	.9212	.8622	.9125	.9604	.9794	.8618
5	.9933	.9898	0.9832	.9681	.9300	.8766	.9222	.9649	.9819	.8763
6	.9939	.9907	0.9847	.9709	.9359	.8865	.9287	.9680	.9835	.8862
7	.9943	.9914	0.9857	.9729	.9402	.8937	.9334	.9702	.9846	.8934
8	.9946	.9919	0.9865	.9744	.9434	.8992	.9370	.9718	.9855	.8989
9	.9949	.9923	0.9872	.9756	.9459	.9035	.9398	.9731	.9862	.9032
10	.9951	.9926	0.9877	.9766	.9480	.9070	.9420	.9742	.9867	.9067
11	.9952	.9928	0.9881	.9773	.9496	.9098	.9439	.9750	.9871	.9095
12	.9954	.9930	0.9884	.9780	.9510	.9122	.9454	.9757	.9875	.9119

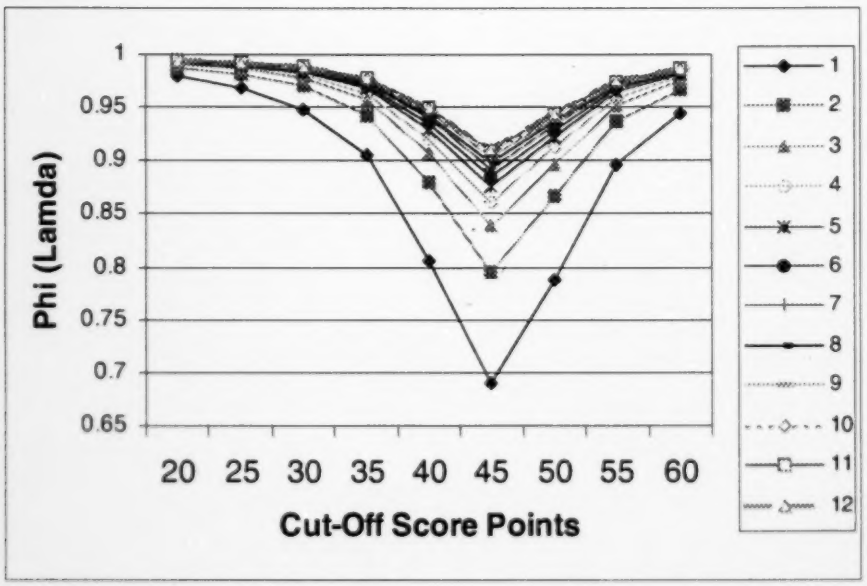

Figure 3. Change of Φ(λ) at different cut-off score points for different test lengths for January 1999 Form (p x R' x I)

No. of Items	Different Cutoff Score Points									
	20	25	30	35	40	45	50	55	60	Mean (45.51)
1	.9766	.9652	.9436	.8978	.7957	.6729	.7659	.8829	.9369	.6712
2	.9865	.9799	.9671	.9392	.8729	.7847	.8524	.9299	.9631	.7834
3	.9899	.9849	.9752	.9538	.9019	.8301	.8855	.9466	.9721	.8290
4	.9916	.9874	.9792	.9613	.9170	.8547	.9029	.9552	.9767	.8537
5	.9926	.9889	.9817	.9658	.9264	.8701	.9137	.9604	.9795	.8692
6	.9933	.9899	.9834	.9688	.9327	.8807	.9210	.9639	.9813	.8799
7	.9937	.9906	.9845	.9710	.9373	.8884	.9263	.9664	.9862	.8876
8	.9941	.9912	.9854	.9727	.9407	.8942	.9304	.9683	.9836	.8935
9	.9944	.9916	.9861	.9740	.9434	.8988	.9335	.9698	.9844	.8981
10	.9946	.9919	.9867	.9750	.9456	.9025	.9360	.971	.9850	.9019
11	.9948	.9922	.9871	.9758	.9474	.9056	.9381	.9720	.9856	.9050
12	.9950	.9924	.9875	.9765	.9489	.9082	.9398	.9728	.9860	.9075

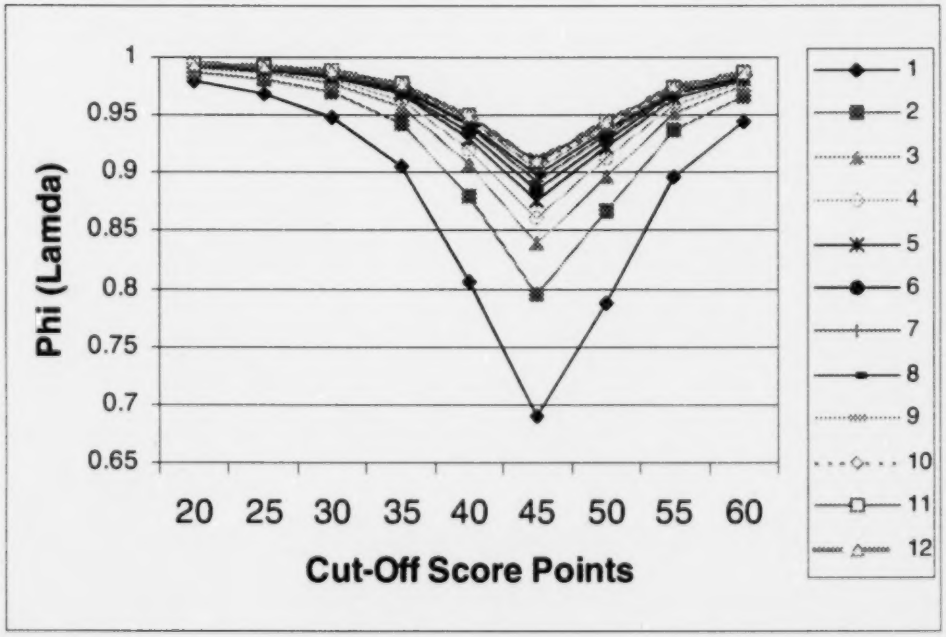

Figure 4. Change of $\Phi(\lambda)$ at different cut-off score points for different test lengths for April 1999 form (p x R' x I)